Frequency-dependent selection in vaccine-associated pneumococcal population dynamics

Jukka Corander, Christophe Fraser, Michael U. Gutmann, Brian Arnold, William P. Hanage, Stephen D. Bentley, Marc Lipsitch, Nicholas J. Croucher

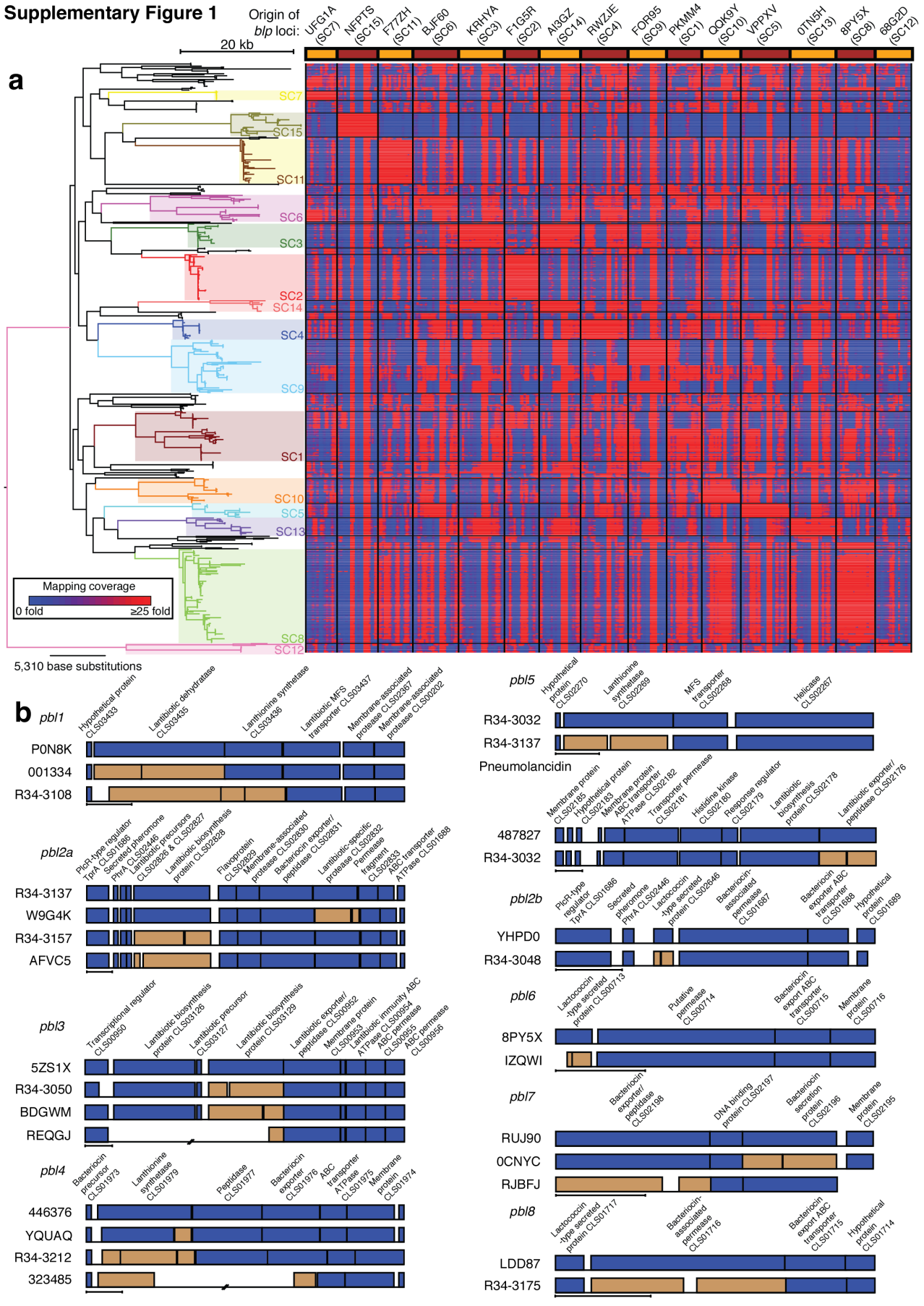
Supplementary Online Materials:

Supplementary Figures 1-10

Supplementary Datasets 1-3

Supplementary Table 1

Supplementary Figure 1



Supplementary Figure 1 Variation in bacteriocin biosynthesis gene clusters in the pneumococcal population.

a Distribution of *blp* (bacteriocin-like peptide) locus alleles across the pneumococcal population sampled from Massachusetts. On the left of the panel is a maximum likelihood phylogeny generated from a core genome alignment. The fifteen monophyletic sequence clusters identified in the original analysis of this dataset are annotated relative to the tree by coloured boxes. The alternating brown and orange boxes at the top of the panel correspond to fifteen alleles of the *blp* locus, spanning the variable region from *blpA* to *blpY*, each extracted from a different sequence cluster. Loci are annotated with the name of the isolate in which they were identified, and the sequence cluster to which the isolate belonged. A 20 kb scale bar is shown to the left of the *blp* loci. The heatmap in the panel shows Illumina sequence read mapping to the *blp* alleles. Each row corresponds to an isolate in the phylogeny, and each column to a base in the *blp* locus allele sequences. Blue regions indicate low mapped sequence read coverage, suggesting the sequence is absent from the corresponding isolate, whereas red regions indicate high coverage, up to a maximum of 25-fold, indicating the sequence is present. The black grid is formed by horizontal lines that demarcate sequence clusters, and vertical lines that separate different *blp* locus alleles. Cells in this grid that are predominantly red represent *blp* loci conserved throughout the sampled members of a sequence cluster.

b Alleles of common pneumococcal bacteriocin-encoding loci (*pbl*) found in the pneumococcal accessory genome. Each set of annotations shows the likely functional form of the locus in the top row, with the full-length coding sequences (CDSs) described and shown as blue boxes. Rows beneath show alleles likely to be deficient in bacteriocin production due to either disruptive mutations to CDSs, which are therefore represented as pseudogenes in brown, or deletions, represented by pairs of diagonal lines across the black horizontal base line. All loci are labelled with

the name of the isolate from which they were extracted. The scale bar underneath each set of alleles represents 1 kb relative to the functional allele.

pbl1: A putatively functional lantibiotic production locus from P0N8K is displayed. Alleles are shown in which bacteriocin production is likely to be blocked by mutations disrupting a CDS encoding one biosynthetic enzyme, in 001334, or genes encoding two biosynthetic enzymes, in R34-3108.

pbl2a: A putatively functional lantibiotic production locus from R34-3137 is displayed. This locus or *pbl2b* may be integrated at a particular site in the pneumococcal chromosome that is associated with the TprA/PhrA PlcR-type quorum sensing system (corresponding to the proteins CLS02446 and CLS01686) characterised by Hoover *et al*. Alleles are shown in which bacteriocin production is likely to be blocked by mutations disrupting a CDS encoding a lantibiotic-specific protease, in W9G4K, or a biosynthetic enzyme, in R34-3157 and AFVC5.

pbl3: A putatively functional lantibiotic production locus from 5ZS1X is displayed. Alleles are shown in which bacteriocin production is likely to be blocked by disruption of a CDS encoding a lantibiotic biosynthesis enzyme by independent mutations in R34-3050 and BDGWM. The same CDS is also removed by a deletion in REQGJ, which additionally eliminates genes encoding a second biosynthetic enzyme and a lantibiotic precursor.

pbl4: A putatively functional lantibiotic production locus from 446376 is displayed. Alleles are shown in which bacteriocin production is likely to be blocked by independent mutations disrupting a CDS encoding a lantibiotic synthetase, in YQUAQ and R34-3212, and by a deletion affecting multiple biosynthetic genes in 323485.

pbl5: A putatively functional lantibiotic production locus from R34-3032 is displayed. An allele from R34-3137 is shown in which a disrupted synthetase gene is likely to block lantibiotic production.

Pneumolancidin: The pneumolancidin production locus from 487827 is displayed. An allele from R34-3032 is shown in which the CDS encoding the lantibiotic processing exporter is disrupted, thereby preventing bacteriocin production.

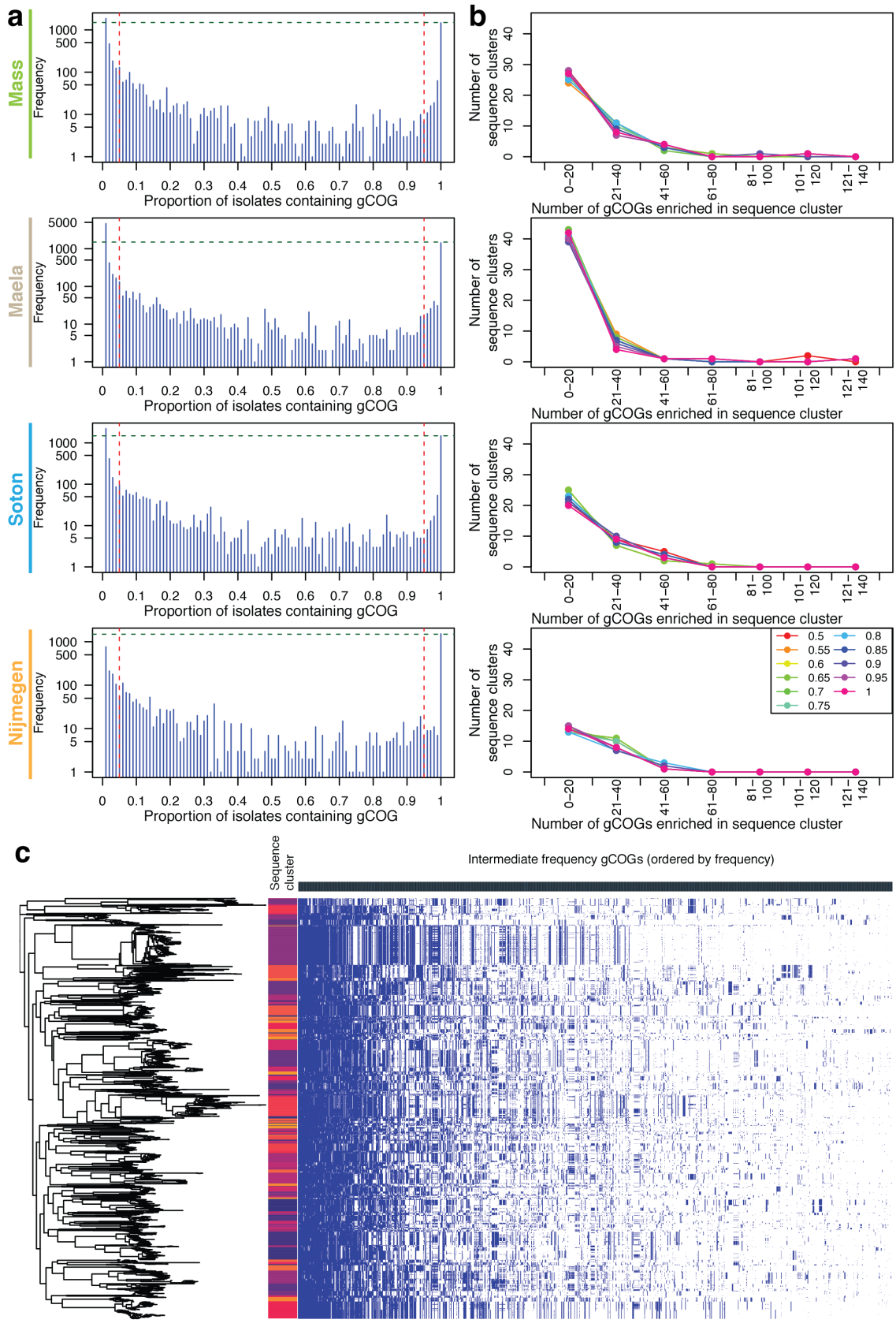
pbl2b: A putatively functional bacteriocin production locus from YHPD0 is displayed. This locus or *pbl2a* may be found integrated at this particular site in the pneumococcal chromosome, which is associated with the TprA/PhrA quorum-sensing system genes. Underneath is shown an allele from R34-3048 in which bacteriocin production may be blocked due to a mutation disrupting the CDS encoding the likely bacteriocin structural peptide.

pbl6: A putatively functional bacteriocin production locus from 8PY5X is displayed. Underneath is shown an allele from IZQWI in which bacteriocin production may be blocked due to a mutation disrupting the CDS encoding the likely bacteriocin structural peptide.

pbl7: A putatively functional bacteriocin production locus from RUJ90 is displayed. Underneath are shown alleles from 0CNYC and RJBFJ in which bacteriocin secretion is likely to be blocked owing to mutations in CDSs encoding exporters.

pbl8: A putatively functional bacteriocin production locus from LDD87 is displayed. An allele from R34-3175 is shown in which a disrupted permease CDS is likely to block bacteriocin secretion.

Supplementary Figure 2

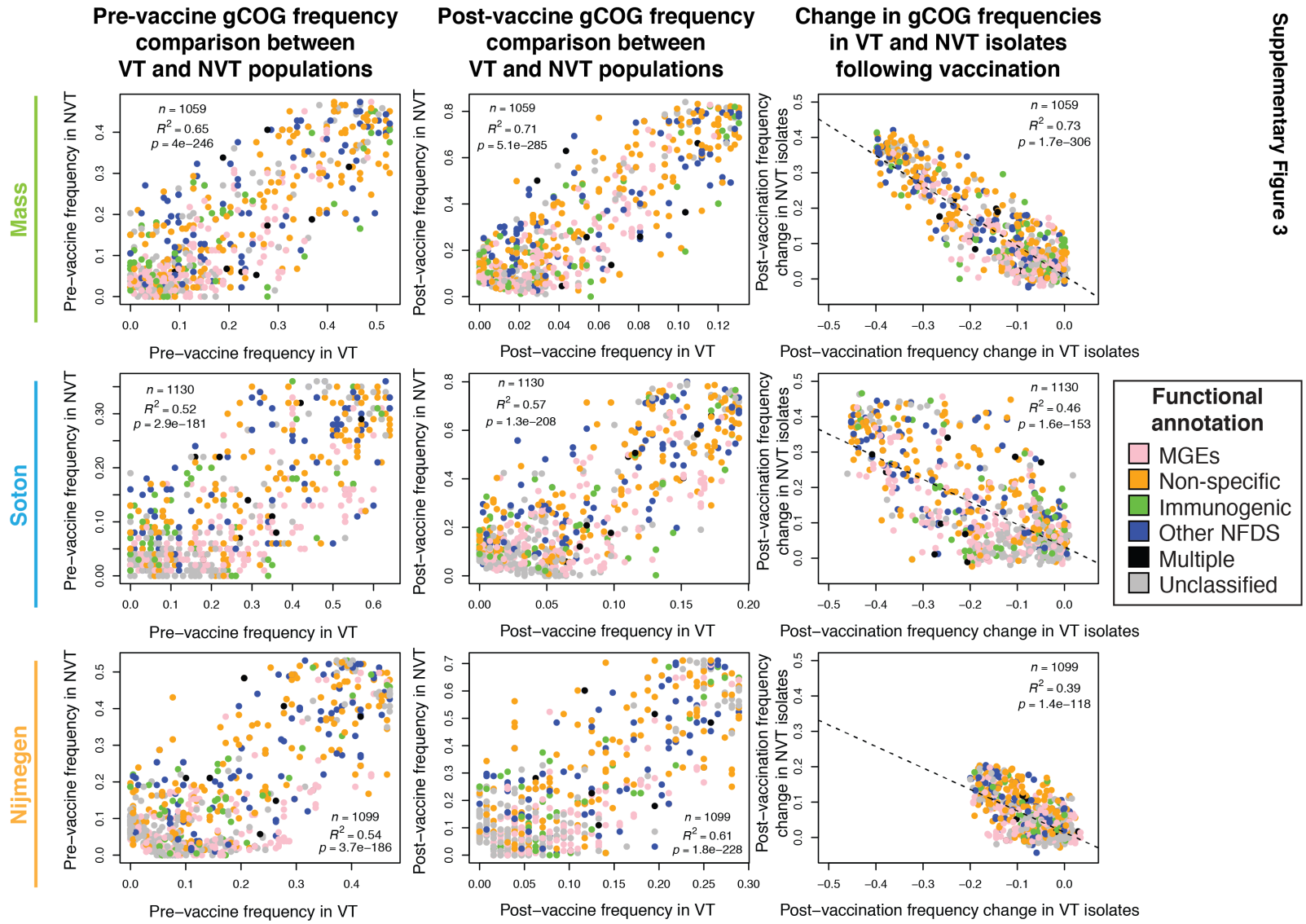


Supplementary Figure 2 Distribution of genes across four pneumococcal populations.

a Each histogram displays the number of gCOGs found in a given proportion of the bacterial population on a logarithmic vertical axis. In each of the four pneumococcal populations, a U-shaped distribution is observed. The size of the 'core' genome is consistently around 1,500 gCOGs, a level indicated by the horizontal green dashed lines. Relatively few gCOGs were found at intermediate frequencies, defined as being present in between 5% and 95% of the population, thresholds marked by vertical red dashed lines. The absolute number of rare gCOGs, present in less than 5% of the population, increases with the number of sampled isolates, being lowest in Nijmegen and highest in Maela.

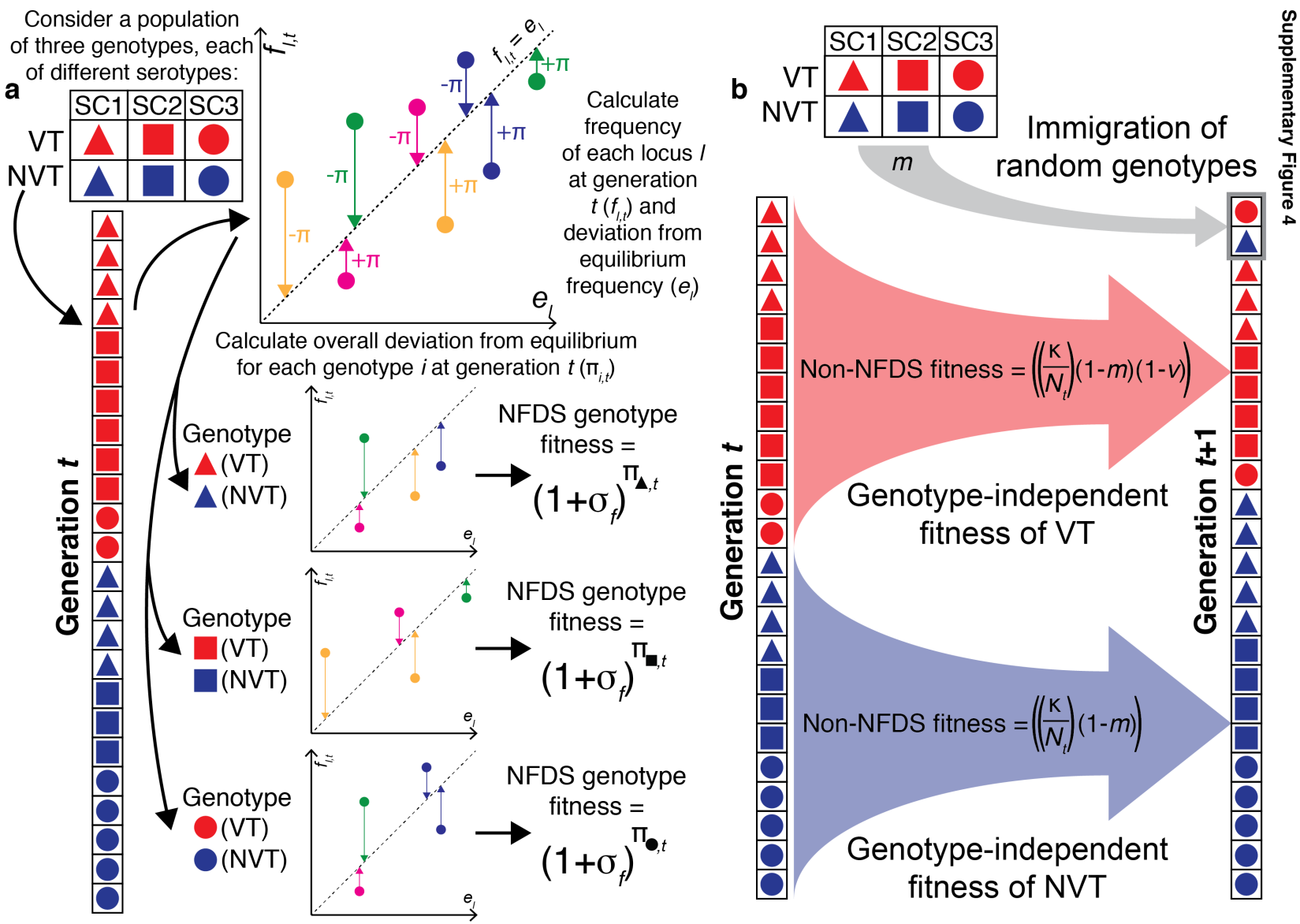
b Distribution of distinctive gCOGs enriched in individual sequence clusters across the four bacterial populations. For thresholds between 0.5 and one, the graphs plot the number of sequence clusters (represented by at least one isolate in the relevant population) with a given range of gCOGs found at, or above, the threshold frequency in that sequence cluster, but no other in the population. The results are qualitatively consistent across all thresholds in all four study populations. The slope is steepest for the Maela population, which is sampled most densely, and therefore is the dataset in which there is the greatest chance of identifying two sequence clusters containing the same gCOG at or above any of the threshold frequencies. Correspondingly, the slope of the curve for Nijmegen, containing the fewest samples, is much shallower. The minor peaks seen in the 61-121 gCOG categories in the Massachusetts and Maela populations represent divergent non-typeable isolates that have many private gCOGs, suggesting they may represent a distinct ecotype or species. However, the majority of sequence clusters are not characterised by much unique genetic content, with many exhibiting almost no distinctive genetic loci relative to other co-circulating sequence clusters.

c Distribution of intermediate frequency gCOGs across the combined pneumococcal population. The core genome phylogeny of the combined genomic dataset is displayed on the left of the panel, adjacent to which is the annotation of the sequence clusters across the combined population, as displayed in Fig 1b. To the right, the columns across the top represent the 1,731 intermediate frequency gCOGs, present in between 5% and 95% of the pre- or peri-vaccination samples of at least one of the studied pneumococcal populations, ordered from the most common to least common. Beneath is a grid in which cells are coloured blue, when a gCOG is present in the isolate at the corresponding position in the tree, or white, when the gCOG is absent from this isolate. The vertical stripes in this grid generally correspond to the extent of sequence clusters, representing the polyclonal pattern of these gCOGs' distribution across the bacterial population.



Supplementary Figure 3

Supplementary Figure 3 Changes in gCOG frequency following PCV7 vaccination. Each plot shows data for gCOGs present at an overall frequency of between 5% and 95% in the relevant population: Massachusetts (top row), Southampton (middle row), and Nijmegen (bottom row). The first column compares the division of the overall gCOG frequency between the VT and NVT isolates in the pre-vaccination population; the scale of the horizontal and vertical axes reflects the relative prevalence of VT and NVT isolates in the pre-vaccination population, respectively. Points are coloured according to their functional annotation, as in Fig 2b. The second column represents the analogous information for the post-vaccination population in the same format. The narrowed range of the horizontal axis reflects the reduced proportion of VT isolates in the population. In all cases, there is a significant correlation between the gCOG prevalences in VT and NVT isolates, indicating their gene contents were similar. However, these correlations are substantially weaker than those between the overall pre- and post-vaccination populations (Fig 2d). This can be attributed to gCOGs rising in frequency in the expanding NVT population to the requisite extent to compensate for their decrease in the declining VT population. This pattern is shown in the third column, which demonstrates the significant anticorrelation between gCOGs' change in overall frequency in the VT and NVT populations between the pre- and post-vaccination samples. The linear relationships between the plotted quantities are represented by the dashed lines and the Pearson correlation statistics shown on each panel, including two-sided *p* values.



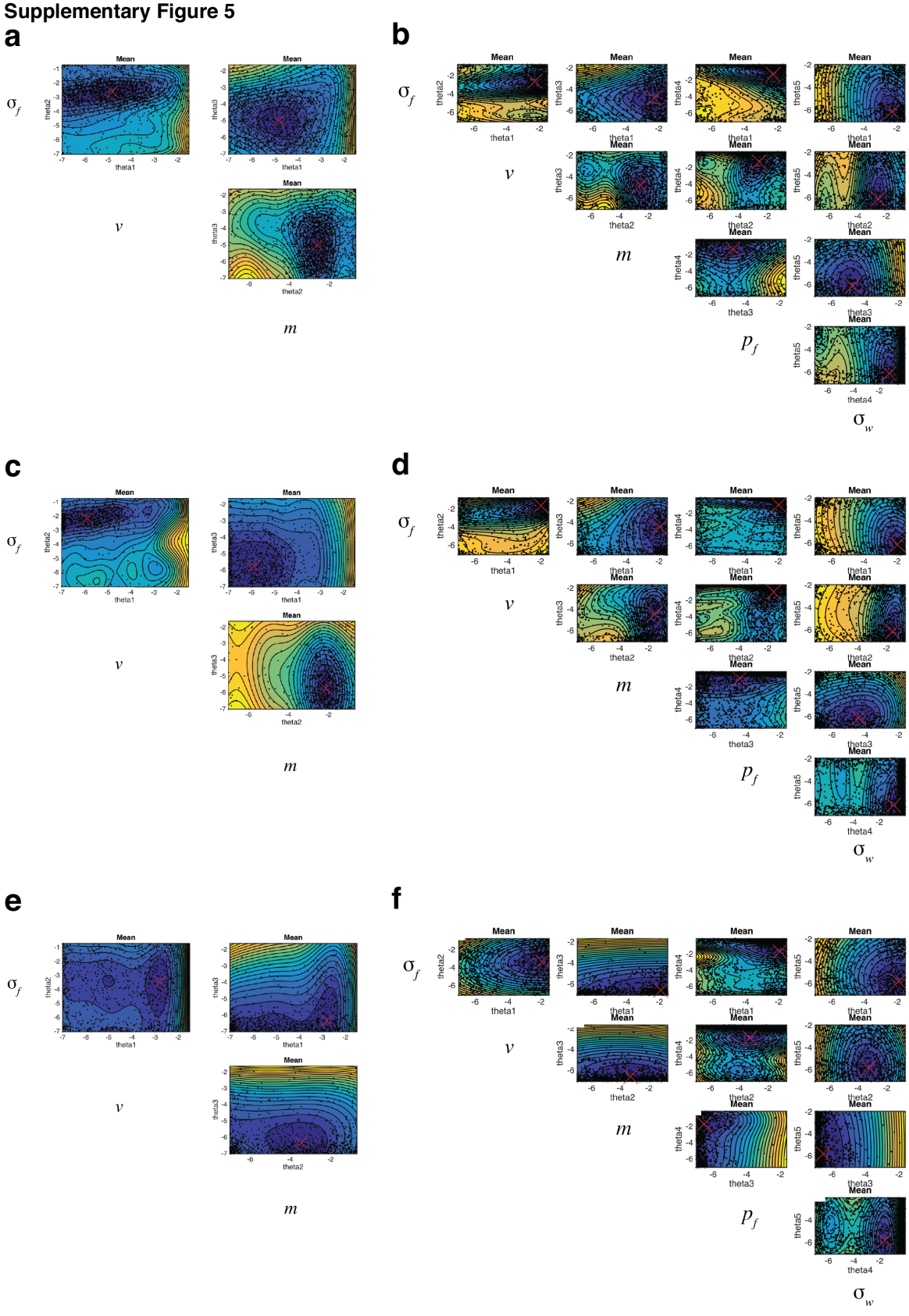
Supplementary Figure 4

Supplementary Figure 4 Description of the multilocus negative frequency-dependent selection model.

a Calculation of genotype fitness. In this panel, the three shapes each represent a particular genotype, with the colour indicating whether these are VT or NVT. At a time *t*, the population will consist of a particular combination of such individuals. From this, the instantaneous frequency of a locus *l*, $f_{l,t}$, can be calculated. When plotted against the equilibrium frequencies of the same loci, e_l, those loci represented by points that lie below the dashed line indicating $f_{l,t} = e_l$ positively contribute to the fitness of genotypes in which they are present (+π), whereas those loci at frequencies above the dashed line negatively contribute to the fitness of genotypes in which they are present (-π). Beneath the overall plot, three smaller graphs exemplify how this same overall distribution results in individual genotypes, with encoding different subsets of the overall *l* loci, having different fitnesses.

b Genotype-independent contributions to individuals' fitness. To maintain an overall stable population size, consistent with the unchanged levels of pneumococcal colonisation post-PCV7, density-dependent selection maintains the overall number of bacteria as approximately κ. This has to account for constant immigration of random genotypes at rate *m*, and therefore the term takes the form $\kappa(1-m)N_t^{-1}$, where N_t is the number of individuals at time *t*. VT individuals suffer an additional cost of (1-*v*), where *v* represents the reduced transmission of VT bacteria due to vaccine-induced immunity.

Supplementary Figure 5



Supplementary Figure 5 Parameter value estimation through Approximate Bayesian Computation with BOLFI. Each panel summarises the output from 2,000 iterations of the BOLFI algorithm, plotted on a logarithmic scale. At the intersection of each parameter pair, isocontour plots show the distribution of the projected Jensen-Shannon deviations, which reflects the approximate likelihood surface, for parameter value combinations tested during the fitting process. The colour changes from yellow to blue as the deviations decrease. The red cross marks the final point estimates of the displayed pair of parameters.

a Fitting the homogeneous rate multilocus NFDS model to the Massachusetts dataset.

b Fitting the heterogeneous rate multilocus NFDS model to the Massachusetts dataset.

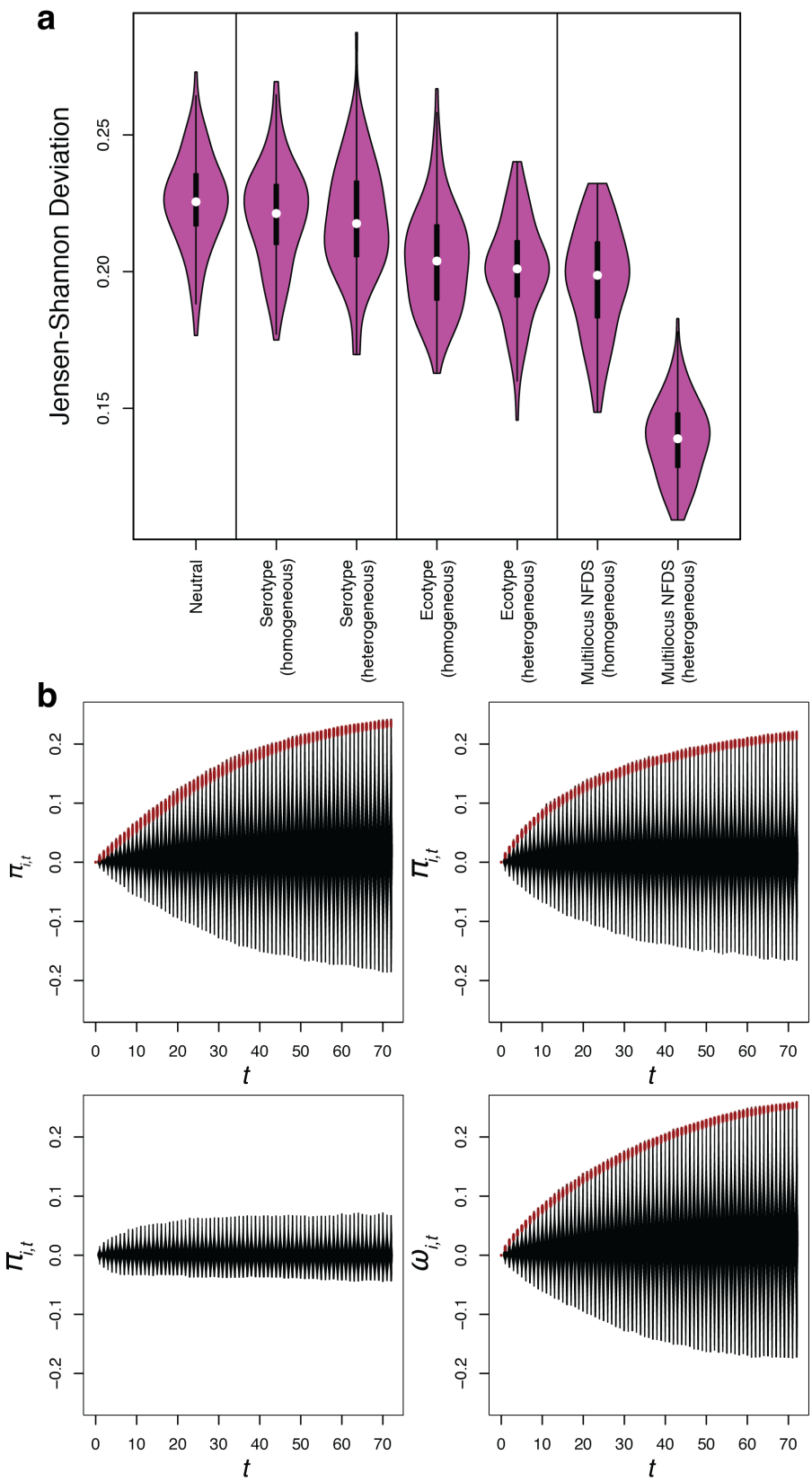
c Fitting the homogeneous rate multilocus NFDS model to the Southampton dataset.

d Fitting the heterogeneous rate multilocus NFDS model to the Southampton dataset.

e Fitting the homogeneous rate multilocus NFDS model to the Nijmegen dataset.

f Fitting the heterogeneous rate multilocus NFDS model to the Nijmegen dataset.

Supplementary Figure 6



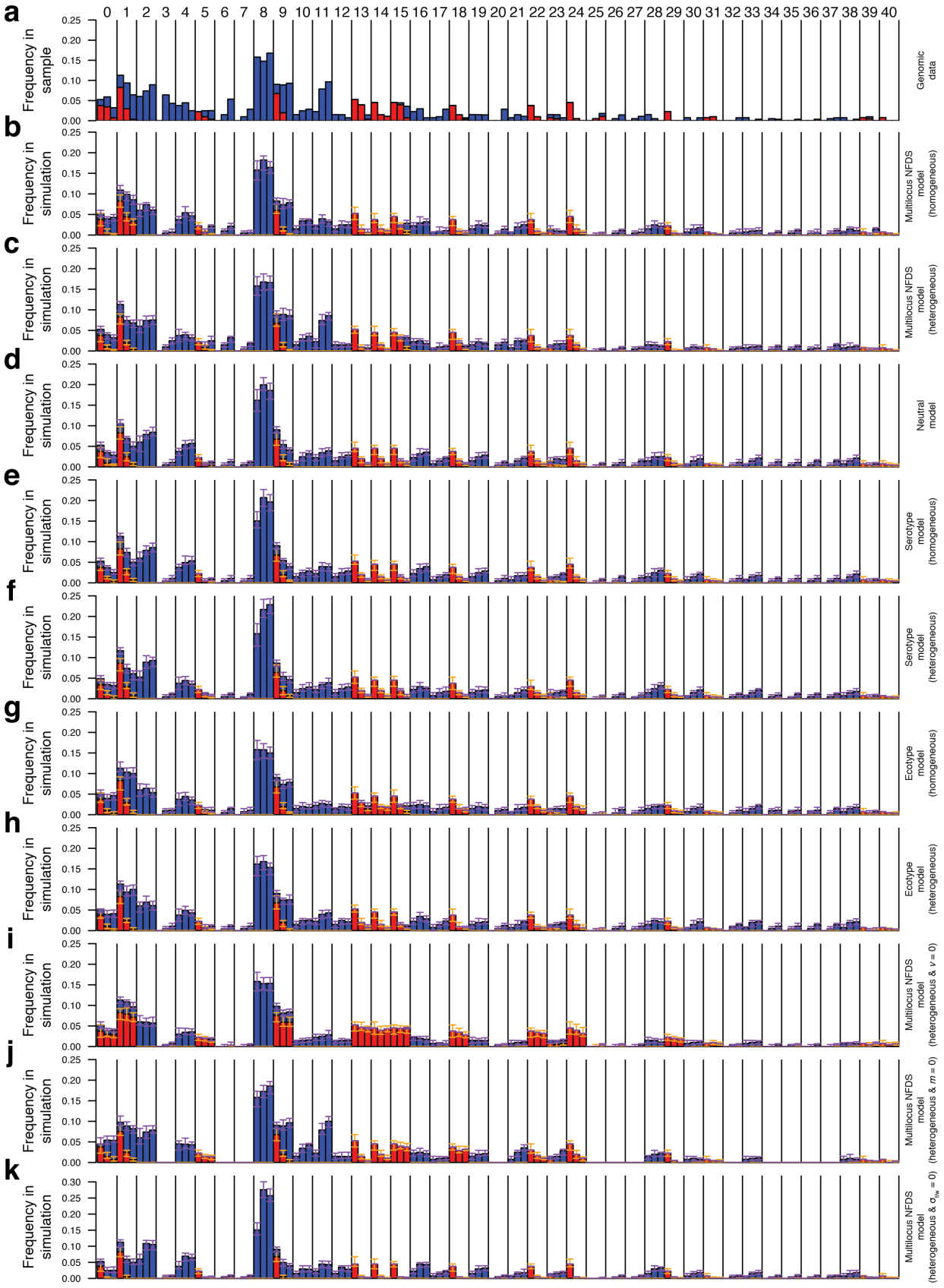
Supplementary Figure 6 Comparison of different model structures.

a All models were fitted through Approximate Bayesian Computation, minimising the Jensen-Shannon deviation (JSD) between the population structures from the Massachusetts genomic samples and simulations (see Methods). These violin plots show the distribution of JSD values from 100 simulations conducted with the point estimate parameter values for each model structure, as shown in Table 1. The two parameter neutral model has the highest median JSD value, indicating it is the worst fit to the data. The serotype NFDS models show little improvement, whether the level of NFDS on different serotypes is homogeneous or heterogeneous. The ecotype models allow for a better fit, but again with no significant improvement in the accuracy of the simulations when the extra two parameters allowing for rate heterogeneity are introduced (comparison of JSD values by Wilcoxon test, $W = 5420$, two-sided $p = 0.305$). By contrast, the homogeneous rate multilocus NFDS model is a significantly better fit than the homogeneous rate ecotype model (Wilcoxon test, $W = 5895$, two-sided $p = 0.0288$), and there is a highly significant further decrease in the median JSD with the addition of two further parameters in the heterogeneous multilocus NFDS model (Wilcoxon test, $W = 9902$, two-sided $p < 2.2 \times 10^{-16}$).

b Deviation of accessory loci from their equilibrium frequencies over each generation during simulations of the Massachusetts population. Each panel displays a violin plot for each simulated generation, reflecting the distribution of $\pi_{i,t}$ or $\omega_{i,t}$, as calculated from 100 simulations run with the appropriate best point estimates of parameter values recorded in Table 1. Red points highlight the distribution of the gCOG corresponding to the *wciN* gene for the synthesis of VT serotypes 6A and 6B, which was previously found to exhibit the greatest deviation from its pre-vaccination frequency in this population based on genomic data (Fig 2). The top left graph shows the distribution of $\pi_{i,t}$ during 100 neutral simulations. The top right graph shows the distribution of $\pi_{i,t}$ during 100 homogeneous rate multilocus NFDS simulations. The

narrower range relative to the neutral simulations shows the constraint imposed by σ_f. The bottom left graph shows the distribution of $\pi_{i,t}$ during 100 heterogeneous rate multilocus NFDS simulations. The narrow range, which reaches equilibrium after around 25 generations, reflects the greater value of σ_f in these simulations relative to the neutral or homogeneous rate NFDS simulations. The bottom right panel shows the distribution of $\omega_{i,t}$ during 100 heterogeneous rate multilocus NFDS simulations. The lower value of σ_w relative to σ_f results in this distribution more closely resembling that of the neutral simulations.

Supplementary Figure 7



Supplementary Figure 7 Simulations of the Massachusetts pneumococcal population. In each barplot, the bacterial population is split into sequence clusters by vertical black lines, annotated at the top of the graph. Each sequence cluster is split into three timepoints: pre-vaccination (2001), a midpoint sample (2004) and a late sample (2007). The bars at each timepoint are split into red segments, for VT isolates, and blue segments, for NVT isolates. Each type of simulation was run 100 times with the point estimate parameter values in Table 1. The bars represent the median values from the combined outputs. The orange error bars represent the inter-quartile range of the simulated VT isolate frequencies, and the purple error bars represent the inter-quartile range of the simulated NVT frequencies.

a The top row shows the sample of sequenced isolates against which simulations were compared.

b This plot summarises the results from the homogeneous rate multilocus NFDS model.

c This plot summarises the results from the heterogeneous rate multilocus NFDS model.

d This plot summarises the results from the neutral model.

e This panel summarises the results from the homogeneous rate serotype NFDS model

f This panel summarises the results from the heterogeneous rate serotype NFDS model.

g This panel summarises the results from the homogeneous rate ecotype model.

h This panel summarises the results from the heterogeneous rate ecotype model

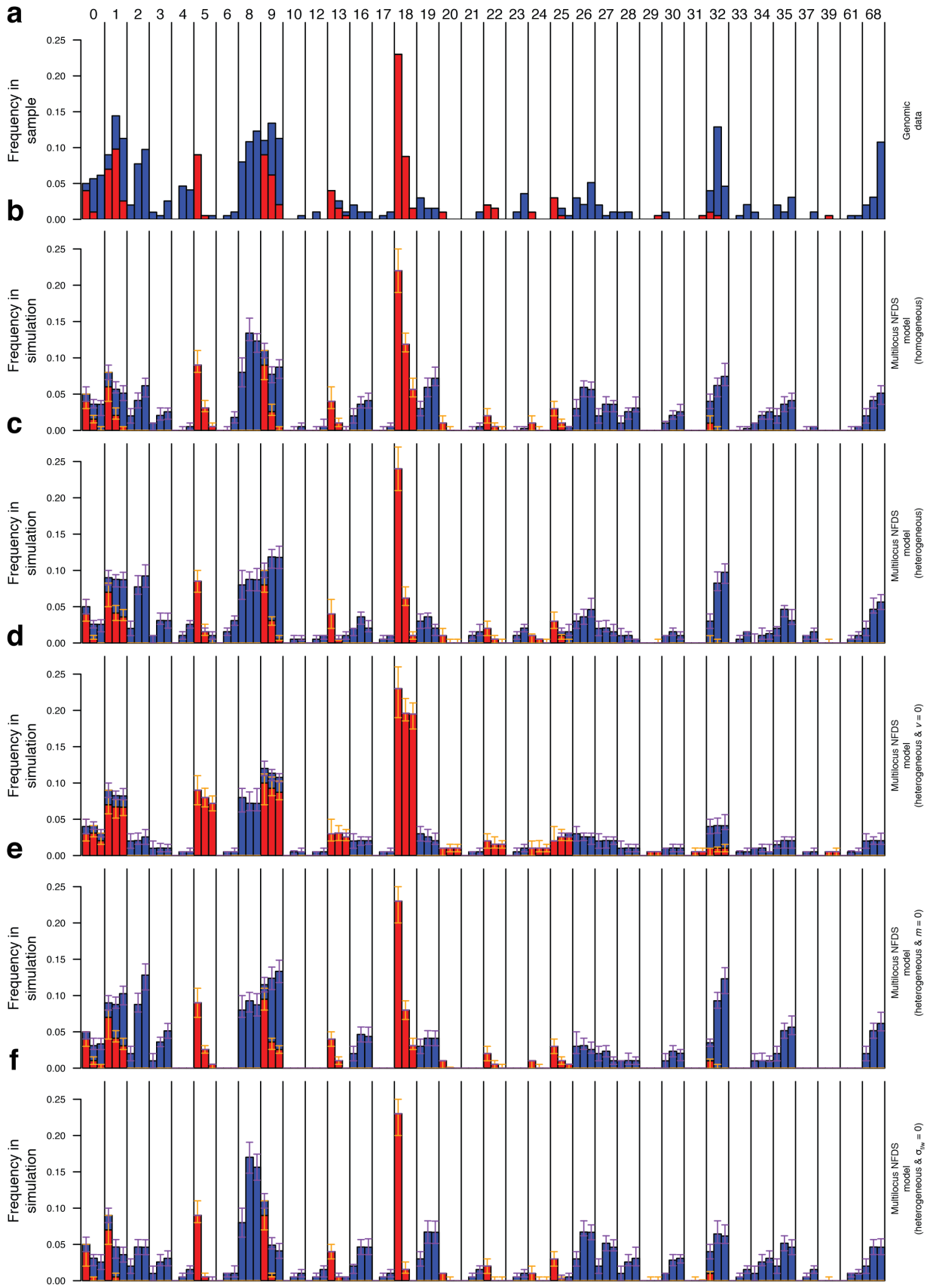
i This plot summarises the results from the heterogeneous rate multilocus NFDS model run with the point estimate parameter values in Table 1, except $v = 0$ to

simulate the absence of vaccination. No substantial decreases in VT isolates' prevalence is observed in this case.

j This plot summarises results from the heterogeneous rate multilocus NFDS model with the point estimate parameter values in Table 1, except $m = 0$ to simulate the absence of migration. While serotype switching within SC1 and SC9 was still observed, the VT isolates within SC5, SC15, SC18 and SC24 reduced in frequency more slowly than in panel c.

k This plot summarises the results from the heterogeneous rate multilocus NFDS model with the point estimate parameter values in Table 1, except $\sigma_f = 0$ and $\sigma_w = 0$ to simulate the absence of NFDS. In these simulations, there is much faster loss of VT isolates, and a much more rapid increase in the frequency of NVT isolates already common in the pre-vaccination sample, such as SC2, SC4 and SC8. The phenomenon of serotype switching is also less evident, as these sequence clusters decrease in prevalence before the switching is complete. These results are very similar to those for the neutral model in panel d.

Supplementary Figure 8



Supplementary Figure 8 Simulations of the Southampton pneumococcal population. One hundred simulations were run with the point estimates of parameter values shown in Table 1. The data are displayed as in Supplementary Figure 7. The bars showing the frequency of each sequence cluster represent three timepoints: pre-vaccination (2007 and before), a midpoint sample (2008-2009) and a late sample (2010-2011).

a The top row shows the sample of sequenced isolates against which simulations were compared.

b This plot summarises the results from the homogeneous rate multilocus NFDS model.

c This plot summarises the results from the heterogeneous rate multilocus NFDS model.

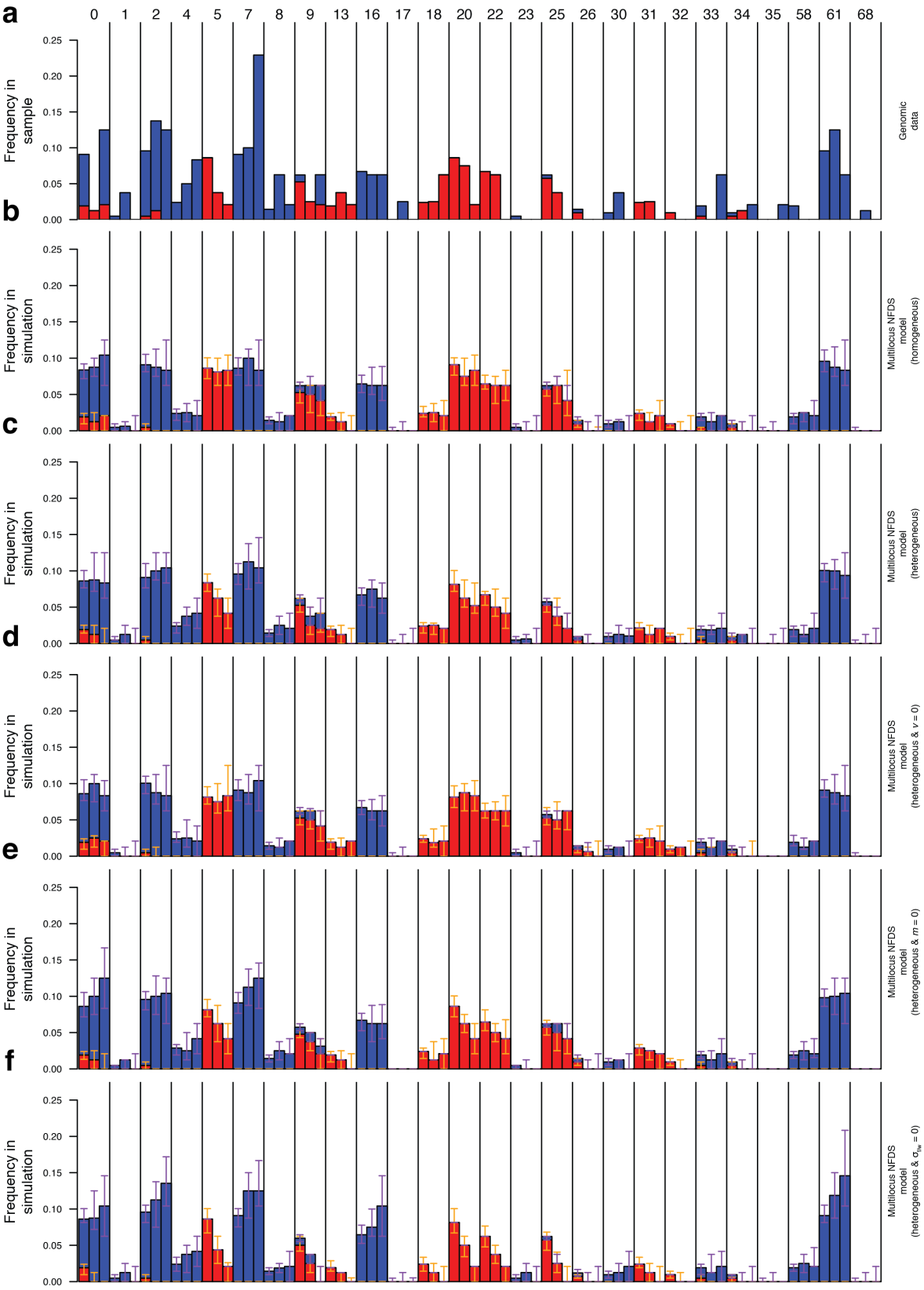
d This plot summarises the results from the heterogeneous rate multilocus NFDS model run with the point estimate parameter values in Table 1, except $v = 0$ to simulate the absence of vaccination. Much less substantial decreases in VT isolates' prevalence is observed in this case.

e This plot summarises the results from the heterogeneous rate multilocus NFDS model with the point estimate parameter values in Table 1, except $m = 0$ to simulate the absence of migration. VT SC18 is cleared from the population less quickly in these simulations.

f This plot summarises the output of 100 simulations using the heterogeneous rate multilocus NFDS model with the point estimate parameter values in Table 1, except $\sigma_f = 0$ and $\sigma_w = 0$ to simulate the absence of NFDS. In these simulations, VT isolates decrease in frequency more quickly than in panel c, and there is less serotype switching owing to VT isolates within sequence clusters being replaced by less

similar isolates before the more similar NVT isolates within the same sequence cluster increase in prevalence.

Supplementary Figure 9



Supplementary Figure 9 Simulations of the Nijmegen pneumococcal population. One hundred simulations were run with the point estimates of parameter values shown in Table 1. The data are displayed as in Supplementary Figure 7. The bars showing the frequency of each sequence cluster represent three timepoints: pre-vaccination (2007 and before), a midpoint sample (2008-2009) and a late sample (2010-2011).

a The top row shows the sample of sequenced isolates against which simulations were compared.

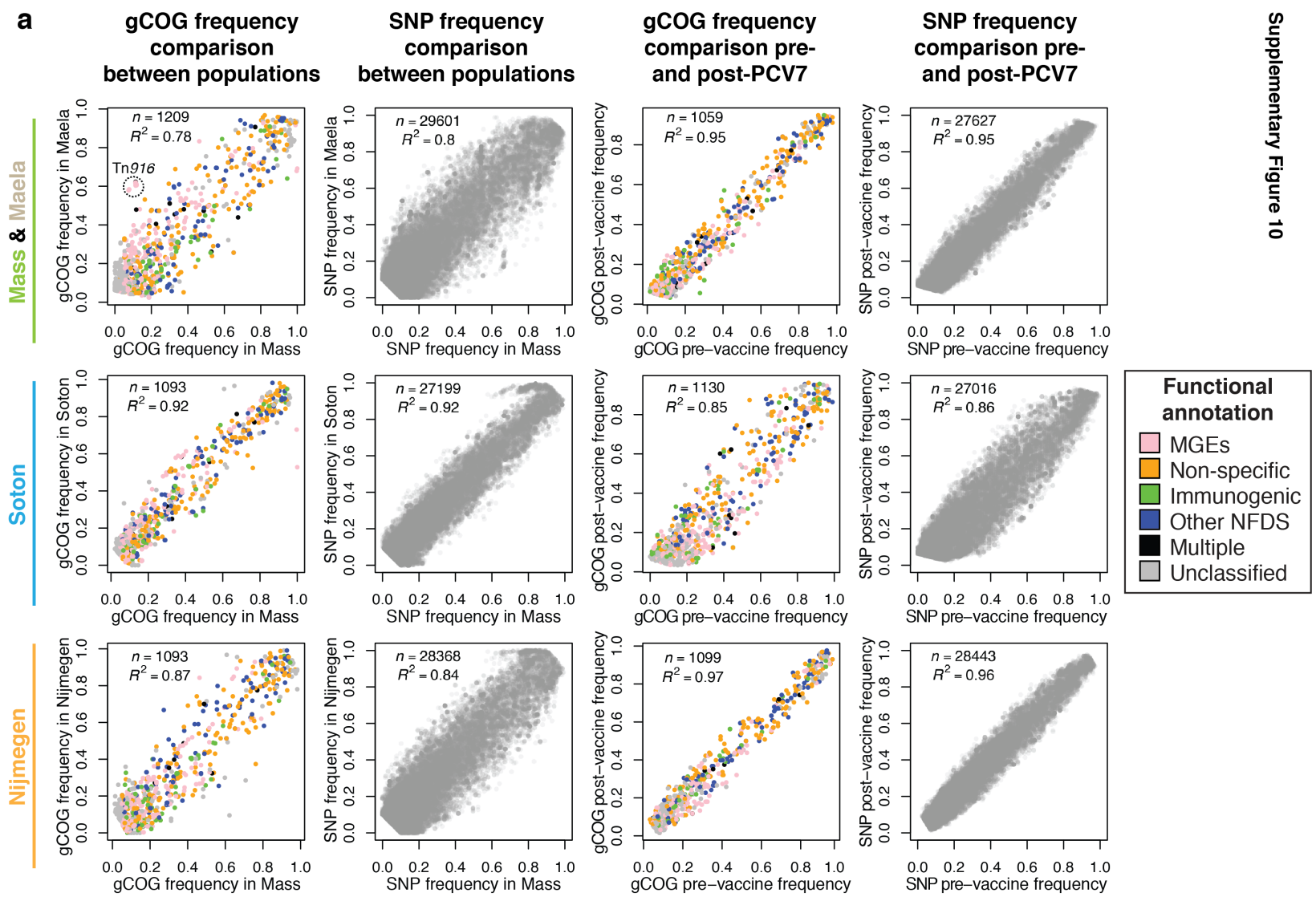
b This plot summarises the results from the homogeneous rate multilocus NFDS model.

c This plot summarises the results from the heterogeneous rate multilocus NFDS model.

d This plot summarises the results from the heterogeneous rate multilocus NFDS model run with the point estimate parameter values in Table 1, except $v = 0$ to simulate the absence of vaccination. No substantial decreases in VT isolates' prevalence is observed in this case.

e This plot summarises the results from the heterogeneous rate multilocus NFDS model with the point estimate parameter values in Table 1, except $m = 0$ to simulate the absence of migration. The output of these simulations show few differences from those in panel c.

f This plot summarises the results from the heterogeneous rate multilocus NFDS model with the point estimate parameter values in Table 1, except $\sigma_f = 0$ and $\sigma_w = 0$ to simulate the absence of NFDS. In these simulations, VT isolates decrease in frequency more quickly than in panel c, and the one instance of partial serotype switching, within SC9, is much less evident, due to the relatively rapid loss of the sequence cluster from the post-vaccination population.

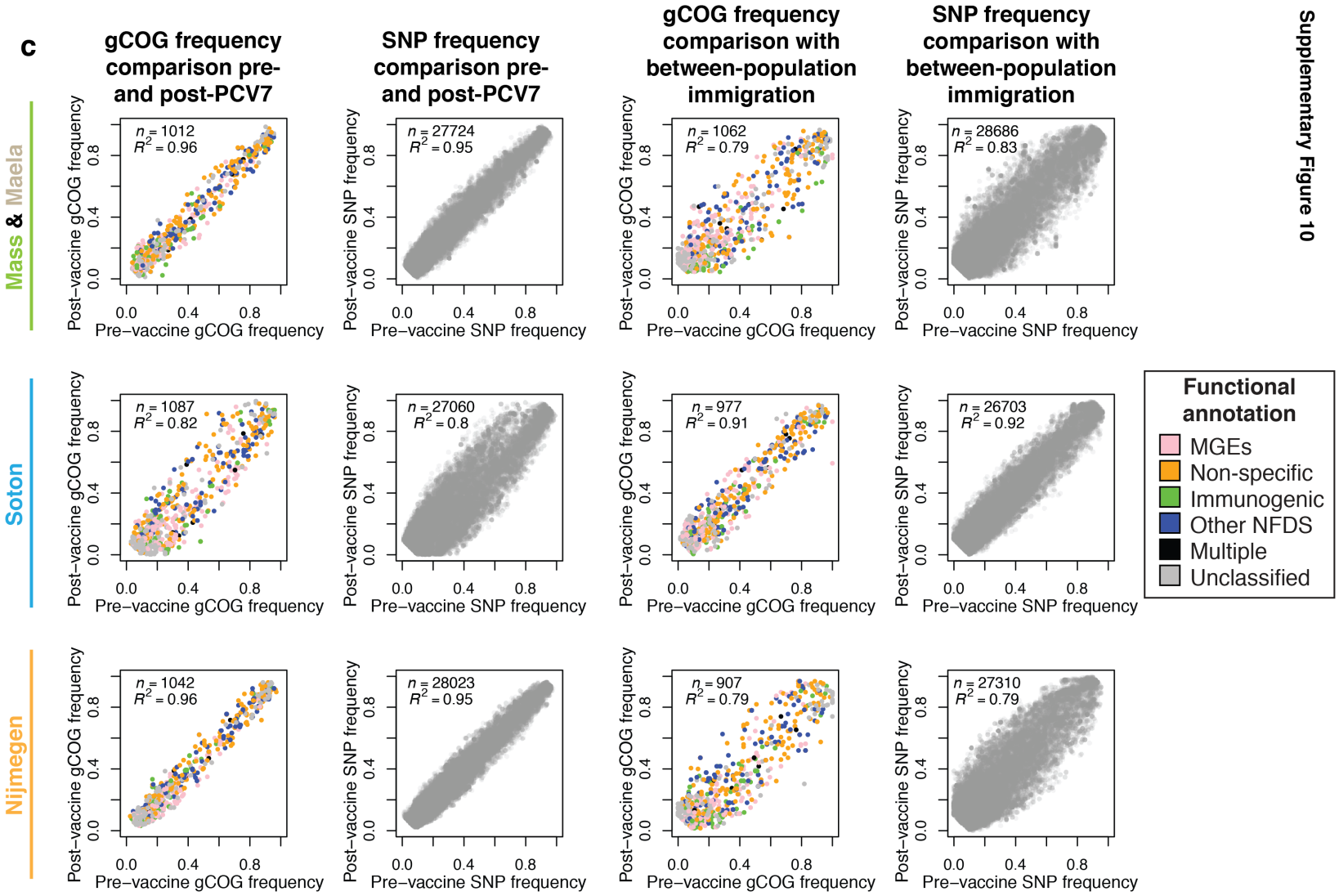


Supplementary Figure 10

Supplementary Figure 10a Comparison of variation in the core and accessory genomes between geographically separate populations, and pre- and post-vaccination within the same population. Loci were included if their overall frequency in the population, or populations, being analysed was between 5% and 95%. As the quantities were linearly related in each case, Pearson correlation statistics are shown on each panel, although the two-sided *p* values are omitted as they were $<10^{-15}$ for each plot. The first column compares the distribution of gCOGs between populations, as in Fig 2b. The frequency of each in Massachusetts is shown on the horizontal axis, and the comparison with frequencies in Maela, Southampton and Nijmegen are shown on the vertical axes from top to bottom. Points were coloured according to the functional annotation in Fig 1a. The second column shows the equivalent correlation between the non-reference allele frequencies of biallelic single nucleotide polymorphisms (SNPs) in the core genome alignment, where the reference allele was that in *S. pneumoniae* ATCC 700669. Although the SNP allele frequencies correlate more strongly than the gCOG frequencies between Massachusetts and Maela, the gCOG frequency correlation is stronger if the Tn*916*-encoded loci were excluded from the compared set of gCOGs (Pearson correlation, $R^2 = 0.84$, two-sided $p < 10^{-15}$; Fig 2). It is likely that the divergence in prevalence of this antibiotic resistance-encoding transposon is a consequence of different selection pressures in the two host populations, hence excluding these loci allows for a fairer comparison of the conservation of allele frequencies between populations. The third column shows the correlation between gCOG frequencies in the pre- and post-vaccination populations in Massachusetts, Southampton and Nijmegen, as in Fig 2d. Loci were only included if their overall frequency in each population across all timepoints was between 5% and 95%. The fourth column shows the equivalent comparison of non-reference SNP allele frequencies, which have a similar correlation to that of the gCOGs in each case.

Supplementary Figure 10

Supplementary Figure 10b Linkage disequilibrium in pneumococcal populations. Using the set of biallelic core SNPs and accessory gCOGs with a minor allele frequency ≥5%, each barplot shows the proportion of pairwise comparisons between loci within the specified separation distance range with an $r^2 \geq 0.4$. Distances were calculated relative to the separation in the *S. pneumoniae* ATCC 700669 genome; only SNP sites and accessory loci present in this genome were included in the analysis. The four rows show analyses of the Massachusetts, Maela, Southampton and Nijmegen populations. The graphs in the first column show the linkage between biallelic SNPs within core gCOGs. The graphs in the second column show the linkage between accessory gCOGs, using the midpoint of each as the location in the genome for calculating the separation between sites. The third column shows the linkage between accessory gCOGs, again using the midpoint as the location, and core genome polymorphisms. These plots show a loss of linkage disequilibrium signal at separations of greater than one or two kilobases in the core genome, whereas significant linkage disequilibrium is detectable between accessory loci at separations of up to 10 kb, reflecting the organisation of accessory genes into distinct genomic islands that are often around this size. The very low linkage disequilibrium evident between accessory loci and the flanking core polymorphisms indicates selection on a genomic island would not substantially affect the distribution of proximal core genome diversity.



Supplementary Figure 10

Supplementary Figure 10c The simulated distribution of core genome polymorphisms under multilocus NFDS acting on accessory loci. The two columns on the left show the distribution of accessory gCOGs and SNPs from simulations of vaccine introduction in Massachusetts, Southampton and Nijmegen, parameterised according to the point estimates shown in Table 1. Each point represents the median pre- and post-vaccination frequencies from 100 simulations, based on randomly sampling an equivalent number of isolates from the first and last generations as genomes were sequenced in the respective genetic datasets in the pre- and post-vaccine periods. Loci were included in the plots if the mean of these median frequencies was between 5% and 95% across both the pre- and post-vaccine samples from the simulations. Despite heterogeneous multilocus NFDS acting only on the accessory loci, and the lack of a midpoint sample meaning the plots are not exactly comparable with those in Supplementary Fig 10a, the simulated core polymorphism allele frequencies closely mirror the equivalent trends in the genomic data shown in Fig 2. The two columns on the right show similar data for simulations in which isolates from different populations were mixed. The pre-vaccine population was that of the Massachusetts dataset, but migration allowed the entry of both these isolates, and isolates from the population of another dataset. Each plot shows the correlation between a sample of randomly selected pre-vaccination isolates, of the same size as the pre-vaccination sample from Massachusetts, and a sample of randomly selected isolates from the final generation, selected to be of the same size as the post-vaccination samples from the non-Massachusetts population. These plots replicate the comparisons of genomic data shown in Supplementary Fig 10a. In the top row, the simulation was parameterised according to the point estimates for Massachusetts, but the e_l values and post-vaccine isolates were from Maela. The consequent post-vaccine divergence replicates the higher correlation between core genome allele frequencies than between accessory loci, despite NFDS acting on the

latter. The second and third rows show the output of simulations in which the population sizes, parameterisation, e_l values and post-vaccine isolates came from the Southampton and Nijmegen datasets, respectively. In these simulations, the correlation between the accessory gCOG and core SNP frequencies are very similar, mirroring the genomic data, despite NFDS again only acting on the accessory gCOGs. The Pearson correlation statistics shown on each panel quantify the linear relationship between the frequencies being compared; the two-sided *p* values were $<10^{-15}$ in each case, and are not shown.

Supplementary Datasets

Supplementary Dataset 1 Functional annotation and classification of the accessory genome of isolates from Massachusetts. The 1,112 COGs present in between 5% and 95% of isolates from the Massachusetts population are listed and annotated, as well as being defined as members of one or more discrete categories used to generate the graphs in Fig 1a with a 'Y(es)/N(o)' classification. The COGs are those defined previously and made available from http://datadryad.org/resource/doi:10.5061/dryad.t55gq.

Supplementary Dataset 2 Functional annotation and classification of the core genome of isolates from Massachusetts. The 1,194 COGs present in a single copy in each of the isolates from the Massachusetts population are categorised and annotated as in Supplementary Dataset 1. The COGs are those defined previously and made available from http://datadryad.org/resource/doi:10.5061/dryad.t55gq.

Supplementary Dataset 3 Epidemiological data and accession codes for the 4,127 isolates included in the combined genomic analysis. Twenty complete reference genomes were included alongside the isolates from the Massachusetts, Southampton, Nijmegen and Maela datasets.

Supplementary Table

Supplementary Table 1 Point estimates of parameter values for the heterogeneous rate multilocus NFDS models generated based on the Gaussian process minimisers from independent fits to the genomic data, also achieved through running BOLFI for 2,000 iterations, as for the estimates shown in Table 1. The 95% credibility intervals are shown in parentheses.

Population	**Model**	**Maximal NFDS strength, σ_f**	**Vaccine selection strength, v**	**Migration rate, m**	**Proportion of loci under strong NFDS, p_f**	**Weaker NFDS strength, σ_w**
Mass	Heterogenous rate multilocus NFDS	0.1017 (0.0185 - 0.2119)	0.0776 (0.0443 - 0.1484)	0.0086 (0.0015 - 0.0209)	0.2916 (0.1178 - 0.5567)	0.0023 (0.0010 - 0.0341)
Soton	Heterogenous rate multilocus NFDS	0.1051 (0.0060-0.1948)	0.1638 (0.0820-0.2756)	0.0025 (0.0010-0.0182)	0.2365 (0.0486-0.4934)	0.0017 (0.0010-0.0462)
Nijmegen	Heterogenous rate multilocus NFDS	0.0618 (0.0011 - 0.1942)	0.0455 (0.0012 - 0.2130)	0.0013 (0.0009 - 0.0057)	0.314 (0.0012 - 0.8199)	0.0111 (0.0010 - 0.1337)